International Conference on Advances in Modern Technology of Research in Engineering Field (AIMTREF) April, 2026
Organized By: SKN Sinhgad Institute of Technology& Science, Kusgaon (Bk), Lonavala, Pune

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Gesture Controlled Music Player Using Hand Gestures

Professor S.P.Gunjal, Pranali Yemale, Jeel Makwana, Tanushree Kshirsagar
Department of Computer Engineering, SKN Sinhgad Institute of Technology & Science, Lonavala, India

Abstract- This paper presents an AI-powered music player that enhances user interaction by integrating gesture recognition and voice-based control with conventional playback mechanisms. The system is developed as a full-stack web application using React for the frontend, Flask for the backend, and MongoDB for data management. It incorporates secure user authentication through JWT and Bcrypt to ensure safe access. The application supports dual music modes, enabling users to play locally uploaded songs stored on Cloudinary or stream music via Spotify integration. A key contribution of this work is the implementation of real-time hand gesture recognition using MediaPipe, allowing users to control playback functions such as play, pause, track navigation, and volume adjustment without physical contact. Additionally, a voice assistant based on the Web Speech API facilitates natural language commands for seamless interaction. The system intelligently switches between local and online sources, ensuring uninterrupted playback. By combining computer vision, speech recognition, and API-based streaming, the proposed solution offers an innovative and user-friendly approach to interactive media systems, highlighting the potential of multimodal interfaces in modern applications.

Keywords- Artificial Intelligence, Computer Vision, Gesture Recognition, Human Computer Interaction, Music Streaming, Voice Control

I. INTRODUCTION

The evolution of Artificial Intelligence (AI) has gradually reshaped how users interact with digital applications, especially in the field of multimedia systems. Conventional music players typically depend on physical interactions such as touch or mouse input, which, although effective, may not always offer the most convenient or intuitive user experience in all situations. As technology advances, there is an increasing expectation for systems to support more natural and seamless forms of interaction.

In this context, gesture recognition and voice-based interfaces have emerged as promising approaches to enhance human–computer interaction. These technologies allow users to interact with systems in a more effortless and contactless manner, improving both accessibility and usability. At the same time, modern users expect flexibility in accessing music, whether through locally stored files or online streaming platforms, making it essential to integrate both capabilities within a single system.

Motivated by these developments, this paper presents an AI-powered music player that combines gesture-based and voice-controlled interaction with traditional playback features. The system enables users to control music using real-time hand gestures as well as natural voice commands, offering a more engaging and user-friendly experience. In addition, it supports both local and online music playback, ensuring flexibility and continuity. By bringing together computer vision, speech recognition,

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

and streaming technologies, the proposed system aims to provide a practical and intuitive solution for next-generation music interaction.

II. LITERATURE SURVEY

In recent years, research in Human Computer Interaction (HCI) has increasingly focused on making systems more natural and user-friendly. One important direction has been the use of gesture recognition, where computer vision techniques are applied to detect and interpret hand movements in real time. Frameworks like MediaPipe have made this process more efficient and accessible, enabling developers to build responsive and accurate gesture-based applications without heavy computational requirements.

At the same time, voice-based interaction has become a key area of development. With improvements in speech recognition technologies, users can now control applications using simple voice commands. Tools such as the Web Speech API have made it easier to integrate voice assistants into web-based systems, allowing for more intuitive and hands-free interaction.

In the context of music applications, most existing systems focus either on local music playback or on online streaming platforms. While these solutions are effective, they often lack flexibility and advanced interaction methods. Only a limited number of systems attempt to combine gesture control, voice commands, and dual music sources into a single platform.

Building on these observations, the proposed system brings together gesture recognition, voice assistance, and both local and online playback. This integration aims to provide a smoother, more flexible, and engaging music experience compared to traditional approaches.

III. PROBLEM STATEMENT

Current music players lack natural and hands-free interaction, relying heavily on touch-based controls that are not always convenient or accessible. There is a clear gap in providing real-time, touchless control mechanisms such as gesture and voice within a single system. Additionally, most applications offer limited flexibility by supporting either local playback or online streaming, rather than a seamless combination of both.

Another key challenge lies in ensuring accurate gesture detection, reliable voice command processing, and smooth switching between different music sources without affecting user experience. These limitations highlight the need for a unified, intelligent solution that enables intuitive interaction, flexibility, and real-time responsiveness in music playback systems.

IV. PROPOSED SYSTEM

The proposed system is designed to provide an intelligent and interactive music listening experience by combining Artificial Intelligence with modern web technologies. The aim is to enable touchless and

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

intuitive control of music playback using gesture recognition and voice commands, while maintaining simplicity and ease of use. The application offers a clean and responsive dashboard, allowing users to seamlessly navigate and control music without complex steps. It reduces dependency on manual interaction and enhances overall user convenience.

The system is a web-based application developed for users who seek a flexible and advanced music control experience. It consists of the following main modules:

- Authentication Module – Secure user signup and login using JWT and Bcrypt.
- Local Music Module – Upload and play songs stored via cloud storage.
- Spotify Integration Module – Stream and control online music using Spotify APIs.
- Gesture Control Module – Detect hand gestures for playback operations.
- Voice Assistant Module – Execute music commands using voice input.

The working steps of the system are as follows:

- User logs in and accesses the dashboard.
- User selects music mode (Local or Spotify).
- User performs actions using buttons, gestures, or voice commands.
- Gesture/voice inputs are processed in real time.
- Backend APIs handle playback control and data processing.
- The system provides continuous and seamless music interaction.

V. METHODOLOGY

The proposed system follows a modular and integrated approach to develop an AI-powered music player with multimodal interaction. Initially, the frontend is designed using React to provide a responsive and user-friendly interface, while the backend is developed using Flask to handle API requests and system logic. MongoDB is used for data storage, and secure authentication is implemented using JWT and Bcrypt.

For gesture control, real-time hand tracking is performed using MediaPipe, where predefined gestures are mapped to specific music control actions such as play, pause, and volume adjustment. For voice interaction, the Web Speech API is utilized to capture and process user commands, which are then translated into system actions. The system also integrates dual music modes. Local audio files are uploaded and managed through cloud storage, while Spotify APIs are used to enable online music streaming and control. All user inputs whether from gestures, voice, or manual controls are processed and synchronized to ensure smooth and continuous playback. This methodology ensures efficient interaction, real-time responsiveness, and seamless integration of multiple technologies within a single platform.

VI. SYSTEM OVERVIEW

The proposed AI Music Player is a full-stack web application designed to provide an intelligent and interactive music control experience through multimodal inputs. As illustrated in Figure 1, the system

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

follows a client–server architecture where the user interacts with the application through a React-based frontend interface.

The frontend acts as the central control unit, receiving inputs from three sources: manual UI controls, gesture recognition, and voice commands. The gesture module, implemented using MediaPipe, captures real-time hand movements through the device camera and translates them into control actions. Simultaneously, the voice assistant module, built using the Web Speech API, processes user voice commands for hands-free operation.

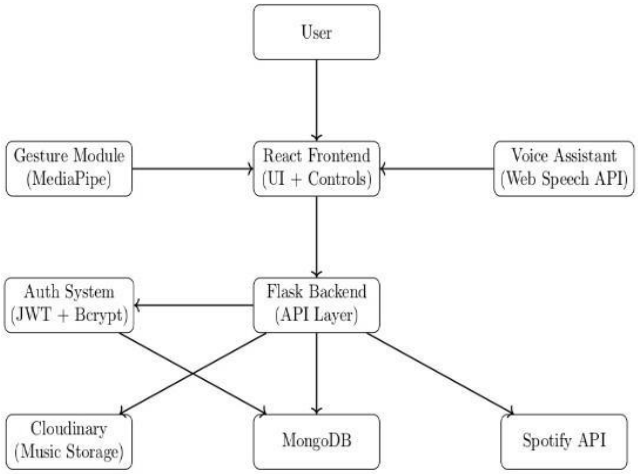


Figure 1. System Architechture

The backend is built using Flask (Python), which functions as the core API layer of the system. It handles incoming requests from the frontend, processes user actions, and coordinates communication between various components. The backend also implements a secure authentication system using JWT (JSON Web Tokens) for session management and Bcrypt for password hashing, ensuring data security and user privacy.

For data management, MongoDB is used as a NoSQL database to store user credentials, preferences, and application-related data. The system also incorporates Cloudinary as a cloud storage solution for managing and streaming locally uploaded music files through secure URLs. In addition, the application integrates the Spotify Web API, which enables users to access and control online music streaming, including features such as play, pause, track navigation, and volume adjustment. OAuth 2.0 authentication is used to securely connect user accounts with Spotify services. The overall system workflow involves capturing user input through multiple channels, processing it in real time, and executing the corresponding music control operations. The backend ensures synchronization between local storage and Spotify services, allowing smooth switching between different playback modes without interrupting the user experience.

By combining technologies such as React, Flask, MongoDB, MediaPipe, Web Speech API, Cloudinary, and Spotify API, the proposed system achieves a robust, scalable, and user-centric architecture. This integration enables efficient real-time interaction, enhances accessibility through touchless controls, and delivers a modern, flexible music playback experience. The model is converted into lightweight TFLite format to ensure smooth, offline performance on mobile devices without relying on internet

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

connectivity. The app can identify various diseases across multiple crops like tomato, potato, maize, apple, and grape, providing both accuracy and accessibility. This system not only minimises the dependency on agricultural experts but also empowers farmers with real-time disease detection, helping them take early preventive actions.

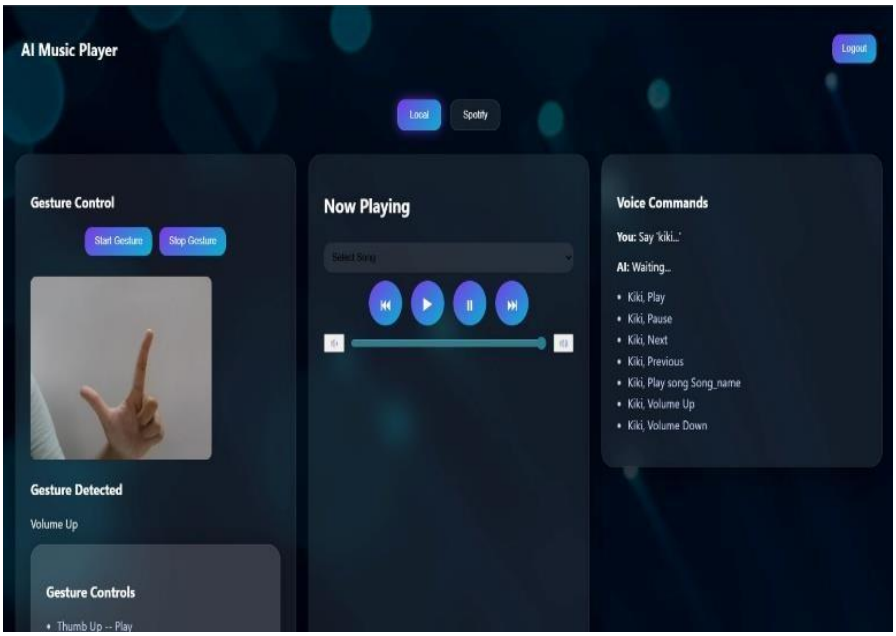
VII. RESULTS

The proposed AI Music Player was evaluated using real-time interaction scenarios involving gesture control, voice commands, and manual inputs. The system demonstrated stable performance and responsive behavior across all modules. Gesture recognition using MediaPipe successfully detected hand gestures such as thumbs up (play) and closed fist (pause), as shown in the results, with accurate mapping to corresponding playback actions.

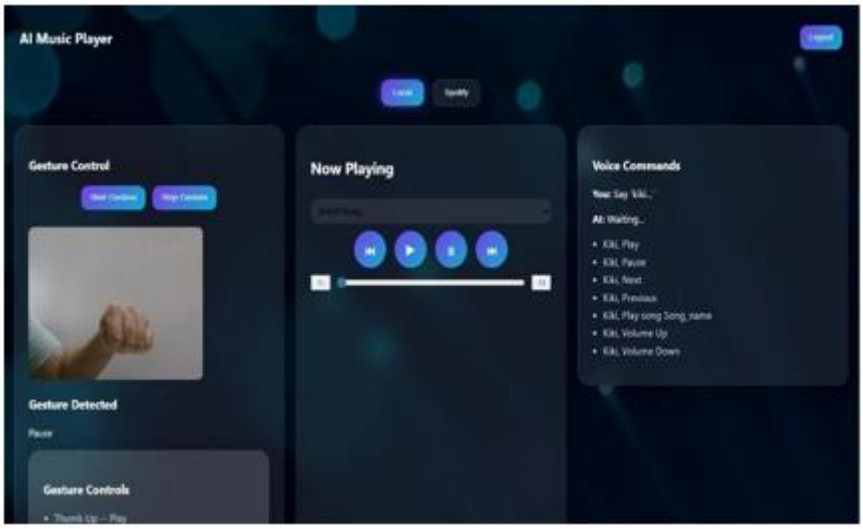
The application interface effectively displays the detected gesture in real time (e.g., “Volume Up” and “Pause”), confirming correct system interpretation. The response time for gesture-based actions was observed to be minimal, enabling smooth and uninterrupted music control.

The UI design is clear and user-friendly, with separate sections for gesture control, music playback, and voice commands. The system allows easy switching between local and Spotify modes, ensuring flexibility in music access. Playback controls, including play, pause, next, and volume adjustment, functioned efficiently across all input methods.

However, minor variations in performance were observed under low lighting conditions or unclear hand positioning, which occasionally affected gesture detection accuracy. Voice recognition may also be influenced by background noise. These limitations can be improved with better environmental conditions and further optimization. Overall, the system demonstrates that integrating gesture recognition, voice control, and web technologies results in an effective, real-time, and interactive music player.



International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752



Regular Music Player	AI Music Player
Touch-based controls	Gesture + Voice control
Requires physical interaction	Fully touchless operation
Single music source	Dual mode (Local + Spotify streaming)
Limited accessibility	High accessibility (hands-free use)
Basic functionality	Smart, interactive system

VIII. CONCLUSION

The proposed AI Music Player successfully demonstrates how multimodal interaction can enhance the user experience in modern music applications. By integrating gesture recognition and voice-based control with traditional playback features, the system provides a more natural, intuitive, and hands-free way of interacting with music. The use of technologies such as React, Flask, MediaPipe, and the Web Speech API ensures real-time responsiveness and smooth performance. Additionally, the inclusion of both local and Spotify-based playback offers flexibility and convenience to users. Overall, the system achieves its objective of creating an intelligent, user-friendly, and efficient music control platform, highlighting the practical potential of combining computer vision and speech technologies in everyday applications.

Future Scope

The system can be further enhanced by improving gesture recognition accuracy using advanced deep learning models and supporting a wider range of gestures. Voice interaction can be expanded to include more natural language understanding and multilingual support for better accessibility. The application can also be extended to mobile platforms for increased usability and portability.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Integration with additional music streaming services and recommendation systems can further personalize the user experience. Moreover, incorporating adaptive AI features, such as learning user preferences and behavior, can make the system more intelligent and context-aware. These improvements can help evolve the system into a more robust and scalable solution for next-generation interactive media applications.

REFERENCES

1. Google, "MediaPipe: Cross-platform ML solutions for gesture recognition," Available: https://mediapipe.dev
2. C. Author et al., "Multimodal Human–Computer Interaction Using Voice and Gesture Recognition," International Journal of Computer Applications, 2021.